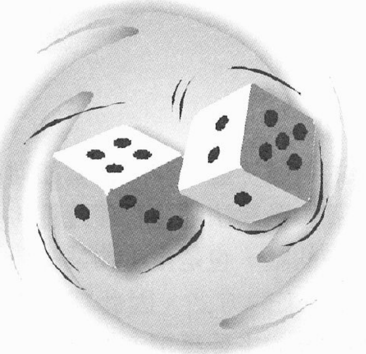


上帝和宇宙玩骰子？－談統計在生物資訊上所扮演的角色

＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊◎農試所農藝組　呂秀英　呂椿棠　魏夢麗

一、前　言

有鑑於生物資訊學在近代基因醫藥及農學研究的重要性，且統計科學在其中具有關鍵性的角色，生物資訊統計的產業競爭已經悄然展開。我們樂見生物學（尤其是分子生物學）和資訊學在這個領域的快速發展，但不幸地是生物資訊相關的統計方法的進展卻十分緩慢。為吸引更多的統計學家加入這個領域，以對現有生物資料的分析提出貢獻，並面臨未來的新型態資料研發出更理想的分析方法，茲將統計在生物資訊上所扮演的角色做一概略的說明。

二、生物資訊統計是生物技術產業的發展關鍵

生物資訊學(bioinformatics)是將數學、統計及資訊技術應用於龐大生物（尤其是遺傳方面）資料之分析與研究的最新竄起之科學領域，興起於人類基因組計畫及其他基因組計畫之產生大量 DNA 資料而蓬勃發展。

根據新華網華盛頓 2002 年 3 月 5 日報導，在過去兩年中，人類基因組草圖的宣佈及基因資訊統計的發展，促使風險投資家和個人投資者將大量資金投入生物技術產業。但專家認為美國生物技術產業的發展步伐可能放慢，其原因除了受到美國經濟景氣衰退、資本市場萎縮等因素影響之外，幾個領域是否有明顯起色是決定性的關鍵：首先，各大研究型公司能否在精練藥物開發方面獲取突破；其次，生物資訊統計領域的研究人員是否能開發出更先進的分析方法及軟體；第三，人類幹細胞研究是否能成為可行性商業的基石，而真正為治療疾病方面帶來好處。

有鑑於生物資訊學在近代基因醫藥及農學研究的重要性，且統計科學在其中具有關鍵性的角色，因此，美國 IBM 等大公司繼續積極投入龐大資金在生物資訊統計研究，而小的軟體開發公司則如春筍般地崛起，成為大公司提供新的研究開發工具的後盾。2001 年年底，美國紐約州撥款 5,000 萬美元支持位於水牛城-尼加拉瓜醫學園的水牛城生物資訊統計研究中心，以研究如何應用超級計算機破解遺傳資訊。如今，美國有一批像水牛城一樣具有發展潛力的生命科學城正等待機會崛起。而我國的國家衛生研究院於 2000 年 11 月也將原成立於 1997 年的生物統計研究組改為「生物統計與生物資訊研究組」，生物資訊統計的產業競爭，已經悄然展開。

三、為何生物學需要統計？

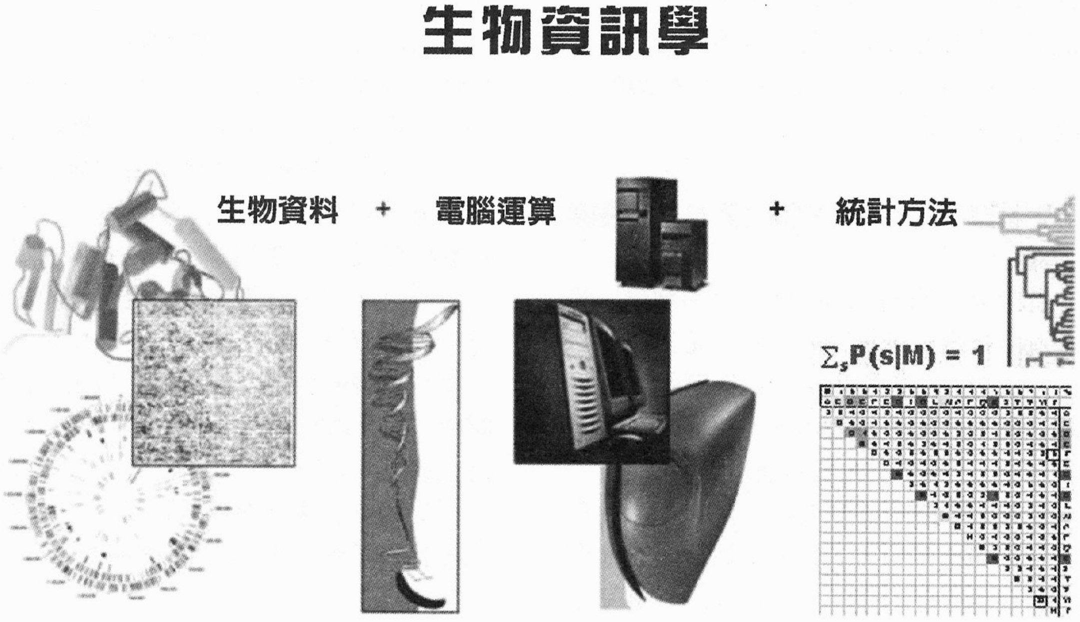
數學和生物是最古老的兩門基礎科學，自人類有文獻記錄以來，這兩門學科即已存

作者：呂研究員秀英
聯絡電話：04-23302301-125

在。然而直到二十世紀數學和生物才有機會結合，並漸漸發展出重要的生物統計學。事實上，許多資料早已存在，經由生物統計學將資料彙整，成為有用的資訊，而透過分析資訊，可歸納出知識。達爾文正是研究了*Galapagos finches*同其他一些物種的形態學特徵，通過比較分析，從而提出了天擇學說。今天，我們對基因和蛋白質序列進行比較，從本質上來說是同達爾文一樣，進行同樣的分析，只不過更加精細、詳盡。在這個意義上，我們從核酸和胺基酸的層次去分析序列的相同點和不同點，以期能夠推測它們的結構、功能及進化上的聯繫。如何發現資訊的相似性，這些就是統計問題。

生物資訊學的目標，並不在於探索如何控制生物系統的基本數學法則或物理性質。這些法則若是存在的話，也是由生物系統所決定之長期累積下來的方式。因此在現階段，數學在這個領域的主要應用，反而是要創造出能讓研究者用來分析資料的工具，而這類工具皆涉及生物系統的統計模式化。

若從「生物資訊」這個字眼的廣義解釋來看，如圖一所示，生物資訊的內涵是在統計方法建立的基礎下，開發各種資訊技術及軟體用具，來分析生物資料。而統計隱藏其中，就像生物資訊背後的那隻「黑手」，常常會被人忽略了它在生物資訊的重要性。我們必須瞭解，統計分析的方法選擇與結論解



圖一、生物資訊學的內涵

釋，取決於如何使這些公理與現實狀況融爲一體。當所擬定的統計方法不夠精確時，即便電腦運算或搜尋的效能很高，其所分析的結果也會讓人質疑。生物系統的統計模式化一開始都是以最簡單的假設著手。沒有一個模式是完全眞實的，因此隨著資料蒐集愈多，模式也必須配合新發現來做修正，故模式會愈變愈複雜，有時還會出現一些特例或直覺上難以置信的延伸狀況，到最後，模式可能就不再適用了。到此地步，有原創想法的人就會提出完全不同的全新模式，啓動一場科學革命。統計革命就是這種模式轉換的例子。所以說，沒有一個模式是最後的模式。也正因爲如此，生物學根據研究對象，發展出各種不同的學門及領域，從過去大的族群層次，到現在極小的基因層次；而統計學爲探索這些不同的生物對象所表現的種種問題，因而也發展了不同的研究領域，例如對於群落，有相應的生態統計及族群遺傳學；對於個體，有相應的數量遺傳學、基因型與環境之交感模式分析等；在細胞分子層次，則建立遺傳圖譜；在DNA及蛋白質分子研究層次，生物資訊學應運而生。

四、上帝和宇宙玩骰子？

統計量必須是隨機的，是因爲我們認定自然界基本上是隨機的。生物資訊學涉及到生物資料的分析，而這些資料的生成又牽涉到很多的偶發機制，最重要的是生物演化和取樣過程本身就存在有隨機性。當然，有些科學家相信決定論觀點，其中最代表這種看法的，就是愛因斯坦說他不相信上帝和宇宙玩骰子的那句名言。以分子生物爲例，儘管表面上是一門倚賴決定論的科學（生物學家發現，基因會使細胞產生特定的蛋白質），但這門科學產生出來的實際數據還是充滿隨機性，而且基因事實上就是這些數據的分布參數(parameter of distribution)。此外，現代藥物對身體功能的藥效，似乎是百分之百確定的，有時一或二毫克就會對疾病產生很大影響，但是證明這些藥效的藥物學研究，卻是依據機率分布(probability distribution)來設計與分析的，而藥效就是這些分布的參數。同理，其他領域也是。這一大群多變而不確定的數據，經過計算分析，研究結果則是以參數的形式來陳述。因此在進入21世紀，統計革命已經在科學界取得勝利，統計觀點已擊敗了決定論，儘管有缺點，但用途廣泛。

一般生物個體的測量性狀，例如高度、血壓或時間等均屬於連續隨機變數(continuous random variables)，它們總是存在於某個値的連續範圍內。但是在生物資訊統計裡，則以離散隨機變數(discrete random variables)（特別是DNA和蛋白質序列）較爲有關，因此各種統計程序所牽涉到的統計理論，重點在於離散變數，而非連續變數。離散隨機變數是符合隨機性的不連續計數，例如擲兩個骰子的結果，其變數的可能數値爲 2,3,...,12。DNA是由A,G,C,T四個核苷酸(nucleotides)構成之非常長的序列所組成，這些序列是在族群內經過很多世代的演變，如隨機突變，而最終固定下來的。因此，兩個相當不同的序列可能源自於同一祖先。我們希望猜測這兩個序列是否具顯著的相似性，即它們是否來自於同一遠親。這是一個機率問題。

相等機率和獨立性爲機率及統計的中心理念，也是最簡化的假設。通常我們必須假設 A,G,C,T 發生在任何位置上的機率是相等

的，而且所有核苷酸之間互相獨立（即序列之產生爲隨機的），但它未必一定滿足，例如我們從左到右讀取一個DNA序列，若是某位置的核苷酸型態以某種方式與其左側位置的出現型態有關，則核苷酸之間便不存在獨立性。因此，就像密碼學一樣，把不同字母編號，密碼分析師可由重複出現的數字，判斷它代表同一個字母，若訊息夠長，且約略知道不同字母所出現的統計次數，再加上一點好運與猜測，密碼分析師通常可以解出來。但是字母的編號也許會隨時改變，這時就不能把重複出現的字母當成相同的字母，也必須瞭解程式的變換原則，當一種密碼程式變換成另一種時，必定有某種規則可循，所以密碼分析師仍可以從中尋找出統計模式，估計規則性，從而破解密碼。

五、生物資訊的演算觀點

目前在生物資訊這個學科裡所使用到的統計程序，通常在不同部份根據假設及實際狀況，會用到不同的統計理論和分析方法。另外由於生物資訊所涉及的資料量極爲龐大，必須要求在最短時間內完成搜尋及比對，因此在演算觀點上，如何發展更快且有效率的搜尋演算法，也是生物資訊統計的一個重要課題。例如，BLAST是目前最常被用來進行序列相似性的搜尋資料庫的程式，幾乎平均每隔幾秒就有全世界的生物資訊研究人員操作BLAST搜尋法一次。隱藏在這個演算法後面的的統計理論，以隨機遊走(或稱醉漢走路, random walk)爲起始，然後經由序列分析理論(sequential analysis theory)和馬可夫鏈(Markov chains)、隱藏馬可夫模式(hidden Markov models)，最後構成了BLAST演算程序。隨著人類基因組及其他重要物種基因組之草圖即將完成，這些資料的分析，亟需全新而目前不可預料的統計分析，因此BLAST也一直推出最新理論的精密版本，以提高搜尋速度、敏感度和實用性，而未來仍有待更新的進階理論。

不同的搜尋程序、序列資料庫以及不同參數都會對搜尋結果產生影響。但不幸地在多種算法不斷出現的情況下，目前卻還沒有一種工具能完成全部有關的序列功能分析。因此，生物統計資訊研究亟需急起直追！

六、統計在後基因組世代的展望

定序完成只是提供讓人們一窺遺傳密碼全貌的機會，欲瞭解基因、應用基因，最重要的是了解其功能。主宰生長、發育、自動調節、行爲及疾病發生等錯綜複雜、迂迴曲折的現象，大大地受同源基因所編譯之核醣核酸及蛋白質所支配，也受到基因的複雜性及基因與周遭環境動態的交互反應所主宰。單一或少數基因的研究，實難完整解釋複雜的基因交互作用現象，而必須從許多基因的功能著手。因此在基因分析上，20世紀若是基因組世紀，21世紀無庸置疑應爲後基因組世紀(postgenome era)。所謂後基因組世代指的是一個已有整個基因組序列的時代。因此我們不必再將時間浪費在選殖少數基因與決定序列，而可追問DNA中資訊的涵義或是應用序列資訊來解決一些問題，也就是生物科技的應用。

近年來，有三項主要的科技被發展來大規模篩選並監測基因的表現：(1)基因微陣列(cDNA microarray)；(2)基因聯結序列分析法(serial analysis of gene expression)；(3)DNA晶

片(chip)。除此三項整體技術較為成熟並已實際應用之外，蛋白質晶片也被發展來偵測多數基因的功能，唯尚未成熟以應用於基因功能分析上。這些技術可提供成千上萬個基因同時進行分析工作，經由自動化程序及化繁為簡的統計方式，相關的基因或一組基因就有可能被分析出來，對基因功能的研究有快速且深遠的影響。

目前這些多基因分析技術的趨勢傾向於將訊息進行歸類、分析比對，即所謂的叢集分析(cluster analysis)。所利用的方法，從 UPGMA 叢集法 (unweighted pair-group method using arithmetic averages)至 SOM (self-organizing map)等不勝枚舉，但其篩選及分析基因的方法仍不敷使用，尚須不斷繼續探索。從縝密的實驗設計、資料標準化、相似係數之擇定、資料過濾之擇定，以及叢集分析法之擇定，仍有待統計學專家能發揮創意，發展一套大家均適用的方法，以加速此類研究的進展。

七、資料探勘

生命科學領域新出現的「資料探勘(data mining)」也是生物資訊統計公司積極從事的工作。資料探勘或資料採礦，顧名思義，即從紛雜龐大的資料中找出隱藏未知的寶礦出來。根據麻省理工學院(MIT)於 2002 年 1-2 月號的「科技評論」(Technology Review)中指出，資料探勘為未來會改變世界的十大新興科技趨勢之一。生物資訊的龐大資料，尤如一座藏量豐富的礦山，取之不盡，用之不竭。要挖掘出決策資訊，須配合統計分析功能，才能去除雜質，找出一些模擬眞實世界的高純度模式，利用這些模式來描述資料中的特徵及關係。

資料探勘的步驟，首先要理解資料及所進行的工作，並獲取相關知識及技術，據以整合與查核資料，若有錯誤或不一致的資料則予以去除，然後發展模式與假設，實際進行資料探勘工作，最後，測試與檢核所採到的資料，以及解釋與使用資料。簡單歸納來說，分析人員進行資料探勘的五部曲為：(1)抽樣、(2)探索、(3)變換、(4)模式、(5)評估。而這正是知識發現(knowledge discovery)的重要手段！

資料探勘所運用到的技術，分為傳統技術和改良技術。傳統技術以統計分析為代表，尤其資料探勘對象多為變數繁多且筆數龐大的資料，故以高等統計學裡所含括之多變量分析中用來精簡變數者特別常用。在改良技術方面，應用較普遍的有關聯資料庫(relational database)、模糊計算(fuzzy computation)、類神經網路(neural network)、遺傳演算法(genetic algorithm)、決策樹(decision tree)、規則歸納法(rules induction)等。有一點很重要的是，沒有一種資料探勘的工具可以應付所有的要求。對於某一種問題，資料本身的特性會影響所選用的工具，所以可能需要用到許多不同的工具和技術，以從資料中找到最佳的模式。在產學界合作下，近兩年來，資料探勘已有驚人的發展，但各種工具只在某些領域下有特別的效能，也就是說尚無適用於所有產業或用途的工具問世。

八、結　語

人類基因圖譜解碼草圖的整個解碼工作預計於 2003 年甚至提前完成，且屆時大部分的模式生物，如水稻，亦可望完成定序工作。但最重要的是了解其功能，令人期待的後基

因組世紀即將來臨。隨著大量基因資訊的獲得，生物資訊學遂形成重要的研究工具之一，而統計科學在其中具有關鍵性的角色。從基因定序、微陣列、資料探勘等各種應用領域，如何將紛雜龐大資料整理、處理、分析和解釋，過濾出有用的資訊，以獲取知識，均涉及生物系統的統計模式化。但至今尚無任何一個工具，可以應付所有的要求。在生物資訊的領域裡，另外一場我們尚未發展的生物資訊統計的科學革命正蓄勢待發。

九、參考文獻

Baldi, P. and S. Brunak. 2000. Bioinformatics: The Machine Learning Approach. forth ed., Cambridge, Massachusetts, London.

Davids, S. 2001. The Lady Testing Tea: How Statistics Revolutionized Science in the Twentieth Century. W. H. Freeman and Company, Inc., New York.

Durbin, R., S. Eddy, A. Krogh and G. Mitchison. 1998. Biological Sequence Analysis: Probabilistic Models of Protein and Nucleic Acids. Cambridge Univ., Cambridge.

Ewens, W. J. and G. R. Grant. 2001. Statistical Methods in Bioinformatics: An Introduction. Springer-Verlag, New York.

Golub, T. R., et al. 1999. Science 286: 531-537.

Michalski, R. S. and K. A. Kaufman. 1997. Data Mining and Knowledge Discovery: A Review of Issues and a Multistrategy Approach, Machine Learning and Data Mining: Methods and Applications. John & Sons Ltd.

Nowak, R. 1990. Science 270: 368-371.

Romesburg, H. C. 1990. Cluster Analysis for Researchers. Krieger Publishing Company, Inc., Florida.

Tamayo, P., et al. 1999. Proc. Natl. Acad. Sci. USA 96: 2907-2912.

Waterman, M. S. 1995. Introduction to Computational Biology. Chapman & Hall, London.